arXiv:2601.13425v1 [cs.CR] 19 Jan 2026

基于区块链的科学数据完整性系统

Gian Sebastian Mier Bello[1][0009−0003−9276−1017]、A Martínez-Méndez[1, 2][0000−0002−1559−9015]、Carlos J. Barrios H.[1, 3, 4][0000−0002−3227−8651]、Robinson Rivas[5][0000−0001−7900−0764]和 Luis A. Núñez[1, 2][0000−0003−4575−5899]

[1] SC3UIS-CAGE，桑坦德工业大学，布卡拉曼加，哥伦比亚[2] GIRG，桑坦德工业大学，布卡拉曼加，哥伦比亚[3] LIG/INRIA DataMove，格勒诺布尔，法国[4] CITI/INRIA SiNDY，里昂，法国[5] 委内瑞拉中央大学，加拉加斯，委内瑞拉 gian2210073@correo.uis.edu.co、alexander2198160@correo.uis.edu.co、cbarrios@uis.edu.co、robinson.rivas@ciens.ucv.ve、lnunez@uis.edu.co

摘要。 如今，大多数高性能计算（HPC）项目都会根据项目目标，从不同来源获取大量数据。其中一些数据体量极大，因此复制这些数据有时并不现实。另一方面，科学研究要求用于不同目的的数据保持不变，以便不同研究团队能够复现结果、讨论理论并相互验证。本文提出一种新方法，帮助研究团队利用区块链验证此类分布式存储库中的数据完整性。区块链最初是为加密货币开发的，如今已展现出广泛的用途。我们的方案确保：1）安全访问数据管理；2）轻松验证数据完整性；3）在遵循同样严格的完整性策略下，便捷地向数据集中添加新记录。我们基于真实科研合作项目——拉丁美洲巨型天文台（LAGO）项目的一个公共数据集子集，开发并测试了一个原型。

关键词： 区块链 · 高性能计算应用 · 科学协作 · 完整性 · 分布式高性能计算

1 引言

如今，在互联网上发布的几乎所有数据都面临着被篡改或未经授权查看的风险；数据可能在网络的不同层级遭到篡改。在此背景下，区块链应运而生。它最初被构想为一种点对点交易系统，但凭借去中心化、不可篡改性和加密等特性，已证明能够有效确保数字记录的完整性和可追溯性。这项技术最广为人知的应用是比特币。根据 Hirsh & Alman [1], 的观察，比特币无需金融中介即可实现安全的货币交易。尽管如此，他们指出，这项技术的首要目标是建立不可篡改的所有权与来源记录。这一特性对于确保科学数据的完整性，并最终推动科学研究的发展至关重要。

近年来，高性能计算（HPC）已成为获取新的科学知识、推动各学科科学研究发展的关键，正如 Silva等人所指出的那样 [2]。鉴于现代计算能力，如今我们能够以前所未有的真实度和精度执行复杂的模拟。如今，许多应用

都需要大量计算资源，涵盖与自然灾害相关的地震分析、石油和天然气勘探、气候与洪水预测、计算流体动力学（CFD）、脱氧核糖核酸（DNA）测序、电磁方程模拟、能源领域分析、材料科学及其他相关领域 [3]。

应用和资源经过优化与校准，以在执行时实现卓越性能，从而产生大量数据。因此，妥善管理这些资源对于促进研究至关重要。拉丁美洲巨型天文台（LAGO）项目 [4]就是一个相关案例。在这一项目中，数据来自由射电望远镜组成的网络，随后利用拉丁美洲国家的高性能计算（HPC）设施进行处理。为了科研目的，数据保持原样且可供访问至关重要，这样各研究团队便能可靠地协作并讨论研究结果。像 LAGO 项目这样的实例中，数据量庞大，给维护整个数据集的可信副本带来了巨大挑战。为此，所提出的系统采用链下存储模型；只有科学数据的密码学哈希和元数据存储在区块链上，而大型数据集本身仍保留在其原有的分布式存储库中。这种方法能够确保数据完整性，同时避免区块链负载过重。区块链并非为存储大型文件而设计，因此该方法也避免了区块大小的限制。因此，在本研究中，我们提出了一种原型，用于支持 LAGO 项目的测量数据和仿真数据的完整性。

将区块链技术与高性能计算（HPC）结合，是一个新兴领域，可能会显著改变计算资源的使用和共享方式。这种新颖的方法带来了管理和使用方面的挑战，包括可扩展性、共识机制、区块链架构，以及如何在大规模科学基础设施中广泛采用。不过，这种集成在异构系统中的信任、安全性、成本效益和性能方面具有优势（包括多尺度架构中的各个层级，例如边缘计算和雾计算），并能提供去中心化高级计算能力，因此推动了相关研究、开发、部署和实施。

在此背景下，所提出的基于区块链的原型通过解决分布式高性能计算环境中的数据完整性和访问控制等关键挑战，发挥了作用。该原型利用 Hyperledger Fabric 的许可型架构来实现这一目标：该架构仅允许经过身份验证的实体和用户参与，并通过**MSP**使用加密身份（X.509 证书）。访问策略通过链码逻辑强制执行，从而能够细粒度地控制哪些用户可以注册、查询或更新与科学数据集相关的元数据。在可追溯性方面，每笔交易——包括原始数据、仿真输入和处理后的输出的注册——都会不可篡改地记录到账本中，为整个数据生命周期提供可验证且带时间戳的审计追踪。这使研究人员能够重建工作流、验证结果，并在分布式高性能计算基础设施中确认各方的贡献。在防篡改可验证性方面，加密哈希可确保对输入/输出文件或元数据的任何更改都会导致其哈希值与存储的哈希值以及链中的内容区块哈希不匹配，从而表明可能存在完整性受损的情况。借助这项技术，可以确保自研究团队使用数据以来，数据未遭篡改或更改。不过，这项技术本身并不能阻止数据被篡改，只能在其他团队需要检查数据完整性或比较数据来源时，作为数据纯净性的保证。

本文结构如下：第 2 节介绍 LAGO 项目的背景。该项目提供了一个强调透明度、低成本和异构性集成的公开数据集；本节还将介绍为实现我们的方案而修订的最新技术水平及关键特征。第 3 节讨论执行

环境，包括分析并选择最佳部署策略、设计原型、展示概念验证，以及验证和微调已部署的原型。第 4 节介绍最重要的结果，第 5 节则进行讨论。最后，我们在第 6 节总结研究发现，并提出后续工作的建议。

2 背景

本节介绍我们工作的基础：LAGO 计划及其对数据的需求，以及区块链技术的最新进展及其与科学数据管理的关系。

2.1 LAGO

拉丁美洲巨型天文台（LAGO）是一座先进的粒子天体物理观测台，旨在研究空间天气现象、极端天体物理事件以及地面大气辐射。该观测台通过分布式网络运行，网络由部署在31个站点的水切伦科夫探测器（WCD）组成，覆盖从墨西哥到南极洲的广阔纬度范围，海拔则从海平面延伸至5000米，涵盖广泛的地磁刚度截止值以及大气吸收／反应水平。LAGO水切伦科夫探测器结构简单而坚固，并集成了多种设备，可实现时间同步、自主运行、机载数据分析，以及远程控制和自动数据传输 [5] [6]。这些探测器在不同环境和地磁条件下采集大量宇宙射线测量数据，形成规模庞大且独特的科学数据集。

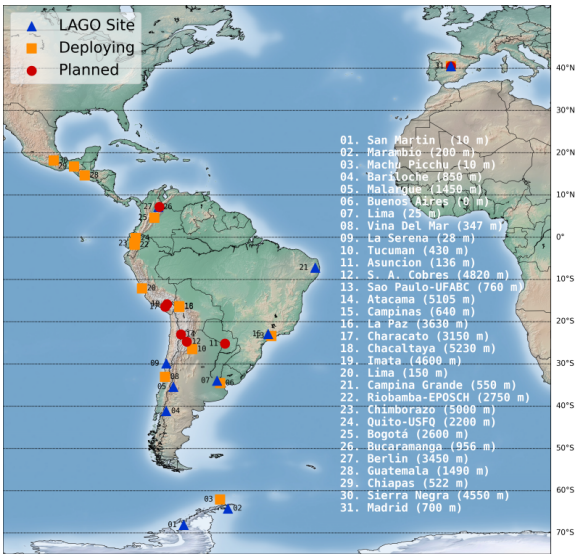


图 1. LAGO 项目 WCD 探测器的地理分布。改编自 [7]

图1展示了探测器所在的不同站点，以蓝色三角形标示。橙色方块表示已部署计算设备并可提供数据存储的站点，红色圆圈则表示计划建设的站点。

LAGO 由 LAGO 合作组织协调，该组织由来自拉丁美洲和西班牙 25 家机构的 90 多名研究人员组成。除了收集实验数据外，该合作组织还使用高性能计算（HPC）框架生成仿真数据。实验数据集和仿真数据集均根据其处理程度和预期科学用途，按分层结构进行组织；较高层的数据量显著减少。

LAGO 的主要目标之一[8] 是保存原始数据和处理后数据，以便在需要时重新处理，尤其是在出现软件错误、分析流程发生变化或方法更新时。鉴于重新生成已损坏或丢失的数据所涉及的计算成本、能耗和复杂性，这一点至关重要。此外，LAGO 广泛使用高性能计算环境来执行模拟和数据分析工作流，因此，确保其输出的完整性和可验证性是实现科学可复现性的基础。在此背景下，本文提出一种基于区块链的解决方案，旨在与高性能计算工作流集成，并支持安全注册数据产品。通过在合作组织范围内提供防篡改、可追溯且可验证的数据集记录，该系统有助于实现更可靠的数据管理，促进跨机构信任，并减少不必要的重复处理工作。

2.2 现状

我们发现了一些相关研究，探讨如何在高性能计算中使用区块链来确保数据完整性并促进科学协作环境的发展。然而，尚未发现能够同时涵盖这些问题的方法。

第一项相关研究是《用于协作科学实验信任的区块链架构》 [12] 。该研究实现了一个由地理位置分散的节点组成的云托管区块链网络，以支持并建立科学协作环境中的信任。该方案利用区块链中记录的不可篡改性来确保数据溯源，并采用去中心化应用，基于溯源模型收集关键数据。其案例研究聚焦于冠状病毒基因组数据，并通过证书进行访问管理，以确保参与者的身份。尽管这篇论文在数据完整性方面与我们的理念一致，但我们着重于支持不同地理区域中的高计算环境。

另一项相关研究是《SciLedger：科学工作流溯源与数据共享平台》 [13]。该方案通过实现区块链来支持科学工作流，具备处理多个数据流和失效机制等功能。它使用默克尔树验证数据溯源，并允许公众访问存储的数据流。虽然我们同样致力于确保数据完整性，但我们的方法有所不同，优先考虑各层对数据流的支持。

最后，在《地球观测数据溯源：一种基于区块链的解决方案》 [14],中，探讨了与地球表面和大气层观测相关的数据的可追溯性与共享。鉴于数据量达到拍字节级，本文提出了一种链下存储模型，利用区块链通过哈希值存储关键信息。此外，本文采用基于投票的共识算法，以优化性能和能效。在这种情况下，如何通过去中心化应用确保数据可追溯、存储数据并保证数据完整性，存在诸多相似之处。总而言之，这些研究展示了区块链

技术增强协作科学的潜力。它们为本提案奠定了基础；然而，它们都没有着重验证数据的完整性。

3 执行环境

本工作基于《区块链实施七阶段》 [9] 开展。具体步骤包括：综述最新技术水平、分析与选择、原型设计、概念验证，以及验证与调整。

3.1 原型架构

在详细研究 LAGO 项目及其在协作中的数据流后，选项被限定为带权限的网络类型。表 1 对本工作所考虑的领先许可型区块链平台进行了比较。

特性	**Hyperledger Fabric [10]**	**Ripple [10]**	**Corda [10]**	**Quorum [10]**
隐私	用于机密交易的私有通道 密交易	基于 信用关系	仅对相关方可见的交易 的	通过 节点和 Tessera 实现私有交易
共识	RAFT	验证节点	由公证人验证	RAFT、IBFT
智能合约	模块化支持 with 链码	不适用	JVM 符合相关法律法规- 定	兼容 Solid- ity 和以太坊虚拟机
自动化	用于业务 规则	仅限于结算	兼容基于法律合同的场景	with 以太坊工具
主要应用一般商业领域、金融,		汇款, 以及全球支付	金融 机构 和法律领域	金融, 物流, and 行业 使用 以太坊

表 1. 区块链平台比较

分析表中的其他替代方案——Ripple、Corda 和 Quorum——后发现，它们都侧重于金融领域。因此，Hyperledger Fabric 2.5 [11] isthe 是 LAGO 项目最合适的选择，因为其权限网络结构可确保只有授权参与者才能访问数据。它还支持私有通道和智能合约，从而提升机密性和运营效率。作为开源平台，它允许用户在无需支付许可费用的情况下进行定制，也便于根据项目需求进行调整。此外，其共识机制 RAFT 即使在发生通信故障时，也会通过选举领导节点来验证交易。这降低了能源和硬件成本，使其实施从长远来看更加可持续。

图 2 所示的区块链原型以 Hyperledger Fabric 为基础平台。研究考虑了 LAGO 合作组织中的两个成员组织：**桑坦德工业大学**和 **钦博拉索高等理工学院**，因为两者都具备测试所需的数据。为支持基于崩溃容错（CFT）的 RAFT 协议，每个组织选择了奇数个节点。下文将详细介绍各组件和交易流程。

1. 证书颁发机构 负责向网络参与者颁发基于 MSP（成员服务提供商）的数字身份。这些身份封装在 X.509 证书中，参与者必须使用这些证书与网络进行交互。

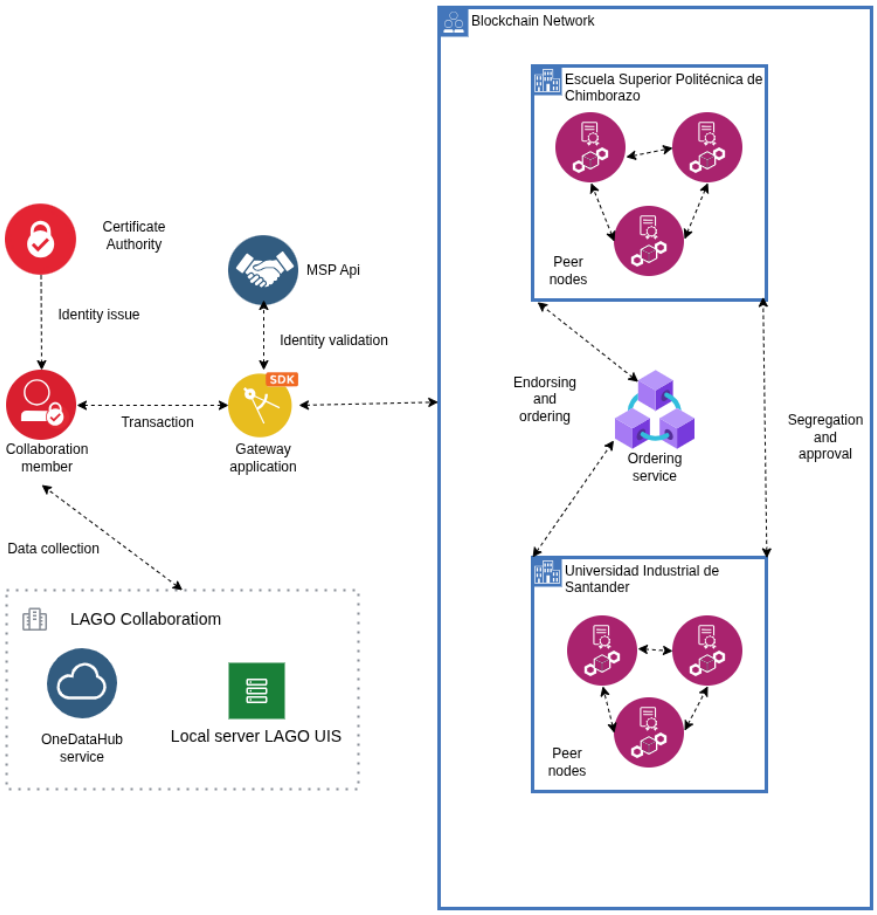


图 2. 区块链原型架构

2. 链码 链码定义区块链网络的业务逻辑和规则，并管理代表系统初始状态和交易的键值对。与传统智能合约不同，链码在节点内的授权私有网络上运行，而非在去中心化虚拟机上运行，并遵循特定的背书策略，以确保安全性和可控性。3. 网关应用程序（软件开发工具包）是用户（协作成员）与区块链网络之间的桥梁。它通过 MSP API 支持提交交易提案和进行身份验证。交易指在节点上运行的链码函数。4. 排序服务 负责接收用户提交的交易，为其分配时间顺序，并将生成的区块分发给两个组织的对等节点。5. 对等节点 每个组织有三个对等节点，它们根据背书策略验证交易，并存储更新后区块链的副本。这确保了记录数据的去中心化。

6. *LAGO* 合作组织 代表了数据收集系统（例如 OneDataHUB 服务和接收测量数据（L0）的本地 LAGO UIS 服务器）与区块链网络的集成。协作成员的测量数据和交易通过交易进行处理和存储。

3.2 数据方案

所提出的数据模式定义了 Hyperledger Fabric 区块链网络中记录信息的通用结构，用于存储 LAGO 项目的科学模拟数据和测量数据。每条记录都有一个**标识符**，它作为唯一标识符，并遵循数据管理计划约定 [8]；还有一个**类型**字段，用于指明记录属于模拟数据还是测量数据及其级别（L0、L1、S0、S1 等）；**元数据**，对应于原始数据的元数据；**原始数据**，包含所获数据的哈希值和位置；**输入数据** 和**输入元数据**，用于存储模拟中使用的输入数据和元数据；**输出数据** 和**输出元数据**，用于记录处理后的结果及其对应的元数据。此外，还包括**站点名称** （生成数据的站点名称）、**合作者名称** （负责合作者的姓名）、**ORC 标识符** （合作者的 ORCID 标识符）以及**访问网址**等附加字段，该字段提供外部地址，以便查询或访问原始文件或相关信息。此结构确保科学工作流程的所有要素都得到清晰定义、可验证且可访问。

3.3 材料

为了在 Hyperledger Fabric 中实现原型，我们使用了一台运行 x86_64 GNU/Linux Debian 6.1.128-1 的机器，配备 7.51 GB 内存和 8 个核心，以及以下工具：**Docker** 和**Docker Compose**。[6] 这些工具对于运行不同的网络服务至关重要，例如包含证书颁发机构（fabric-ca）、对等节点、NoSQL CouchDB 数据库（couchdb）和排序节点镜像的**镜像** 容器。

我们使用 X.509 证书配置了数字身份，以确保对网络的受控访问。我们建立了跨组织通信通道，并部署了“ScientificData-CollectContract”链码，以通过基于角色的访问控制来管理科学记录。

我们发现，如果对等节点因连接中断或其他故障而与网络失去同步，它可能在本地持有一个不同步或过时的账本版本。向此类节点提交的交易可能会在排序和验证阶段被更广泛的网络拒绝，因为这些交易基于不一致的世界状态。这种行为并非协议疏忽，而是一项防止“双重支付”或引入无效数据的基本安全功能。为降低这种风险，我们将客户端应用程序设计为：一旦检测到某个对等节点的账本状态已损坏或不一致，便停止向其提交交易提案，详情见算法 1。

该**网关**应用使用**Node.js**开发，用于与网络交互。它通过比较存储的数据与实时计算的数据，验证区块链中区块的完整性。首先，将内容的哈希值（contentHash）与区块头中记录的哈希值（dataHashString）进行比较。如果两者不匹配，则认为该区块可能已被篡改。然后，提取并验证区块中的交易。

对于区块中的每笔交易，如果其不为空，则提取签名、负载和头部，以及标识符（交易 ID）和发起方的身份。随后准备一个签名验证器

[6] Docker 和 Docker Compose 是用于管理容器的工具：https://www.docker.com/

算法 1 验证区块完整性

过程 完整性验证（contentHash、dataHashString、区块、i）
  如果 contentHash ≠dataHashString **则**
    $blockDataList \leftarrow$ 从区块中获取交易列表。
    如果 $blockDataList$不是 $empty$ **则**
      对于 每个 bl 在 $blockDataList$ **执行**
        如果 bl 不等于 $empty$ **则**
          $signature \leftarrow$ 提取交易签名。
          $envPayload \leftarrow$ 获取交易载荷。
          $headerPayload \leftarrow$ 提取交易头。
          $txId \leftarrow$ 获取交易标识符。
          $creatorBytes \leftarrow$ 提取交易创建者身份。
          $verifier \leftarrow$ 使用 SHA-256 算法准备签名验证。
          使用交易数据验证创建者签名。
          $isValid \leftarrow$ 使用 $creatorBytes$ 对 $envPayload$ 进行签名，并与 $signature$进行比较
          如果 $isValid = false$，**则**
            抛出错误，指出 ID 为 $txId$的交易中的区块 i 已被篡改
          结束条件判断
        else
          打印：区块中没有交易数据
        结束条件判断
      结束循环
    结束条件判断
  结束条件判断
结束过程

并使用 SHA-256 验证创建者签名是否真实。如果验证失败，则抛出错误，指出区块 i 中交易 ID 为该值的交易已遭篡改。如果没有交易数据，则打印一条消息说明其不存在。实现代码库位于 [7]

数据收集 该流程首先从两个主要来源提取部分数据。一方面，选取并下载存储在 LAGO 项目本地服务器中的测量数据；另一方面，从 One Data Hub 的集中式仓库获取模拟数据。One Data Hub 是一个用于存储和访问科学数据的平台。这两个来源提供了待处理的基础信息。

提取的数据使用 Python 脚本进行处理，这些脚本旨在将测量数据和模拟数据转换为标准、轻量级的类 JSON 结构。在此处理过程中，对输入和输出文件及其各自的元数据进行了分析。为确保数据完整性，使用 SHA-256 哈希函数为每个文件及其内容生成唯一标识符，并将其应用于原始格式的数据。此方法可用于验证数据在处理期间或处理后是否未被篡改。

最后，将处理后数据及其转换脚本存储在 GitHub 仓库中。该仓库可通过 [8] 访问，其中包含生成的 JSON 文件以及用于转换数据的代码。这种集中式存储便于访问处理后数据，也使其他协作者能够复现该流程，或根据新的需求进行调整。

随后，将此前的数据存储到区块链中。借助 Hyperledger Fabric 开发的 Web 服务 **Hyper-ledger Explorer** [9]，可以将区块、节点、交易、链码和通道等网络数据可视化。如图 3 所示，进行了 434 笔交易，

[7] 实现代码库：https://github.com/GSMier/ProyectoLago [8] 脚本和处理后数据的代码库链接：https://github.com/GSMier/LAGO-data [9] Hyperledger Explorer 代码库：https://github.com/hyperledger-labs/blockchain-explorer

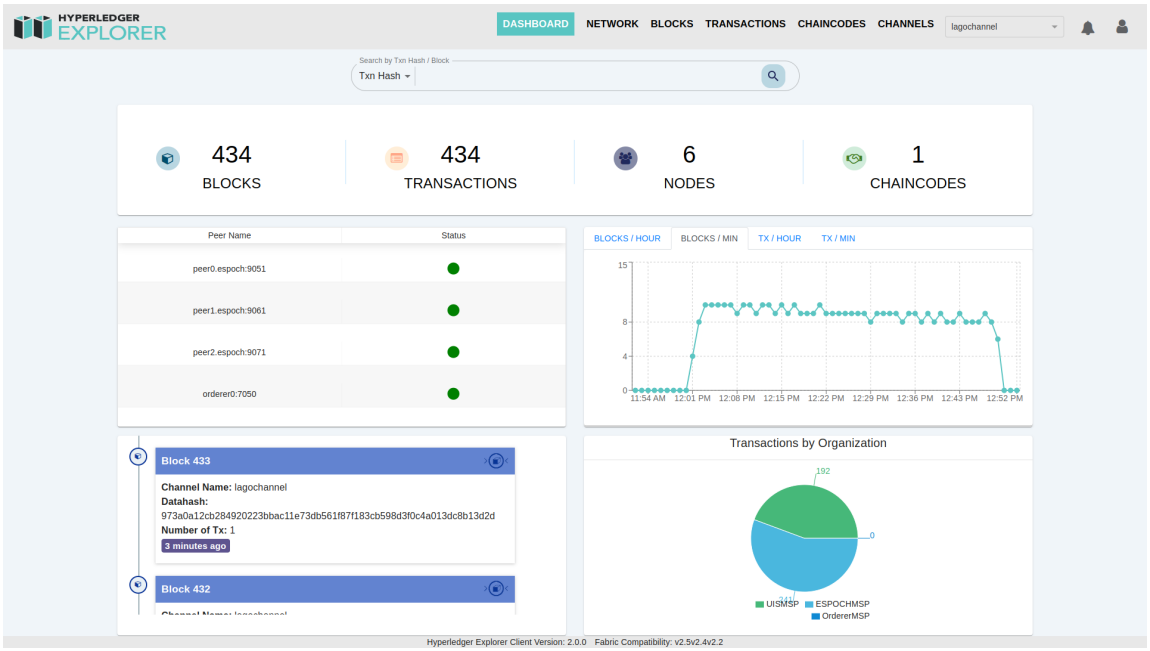


图 3. Hyperledger Explorer 仪表板

其中 430 笔由链码“ScientificDataCollection”发起，用于存储数据。这些数据分布在不同的区块中，因为每个区块以 2 秒为间隔接收交易。两个组织的科学数据以平均每分钟约 8 笔交易的速度被记录下来。在 Hyperledger Fabric 中，其结构更为复杂，因为每笔交易都包含由创建者的私钥生成的数字签名。此外，验证并批准这些交易的对等节点也会将其签名添加到区块中。图 4 展示了这一结构，其中每个区块都通过密码学哈希与前一个区块相连。

3.4 概念验证

区块链是存储网络上所有已记录交易的组件。在 Hyperledger Fabric 中，其结构更为复杂，因为每笔交易都包含由创建者使用其私钥生成的数字签名。此外，验证并批准这些交易的对等节点也会将其签名添加到区块中。

在 Hyperledger Fabric 中，通过验证区块和交易来确保区块链中数据的完整性。在区块层面，通过检查前一个区块的哈希值是否与当前区块中存储的哈希值一致，确保链的连续性，避免历史记录被篡改。同时，还会验证当前区块的哈希值是否与其内容相符，确保区块未被修改。此外，还会验证元数据签名和配置，以确认区块由授权实体发布，且网络配置一致且合法。对于交易，则会检查头部和批准中的数字签名，确保其中涉及的身份属于

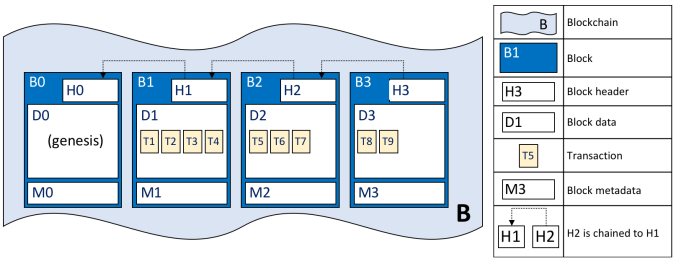


图 4. Hyperledger Fabric 区块链结构——引自 [16]

成员服务提供商（MSP）中的授权实体。还会验证批准者的签名是否合法，且与交易中批准的内容相符，从而确保共识流程得到正确遵循。

在 Hyperledger Fabric 区块链结构中，数据完整性通过特定的密码学机制得到保障。首先，所有区块都通过 **SHA256** 哈希密钥进行密码学链接，每个区块头都包含前一个区块的哈希值，从而形成不可变链[cite: 164,165]。其次，每笔交易都通过**椭圆曲线数字签名算法 (ECDSA)**进行保护。通过**成员服务提供商 (MSP)**，每位参与者都会获得一对公钥/私钥。交易使用创建者的私钥签名，从而确保真实性和完整性，因为任何对交易数据的更改都会使签名失效。为评估数据完整性，测量数据和模拟数据的链上记录（例如数据哈希、元数据和协作者姓名）会直接在存储区块链的文件中修改，并使用由 Hyperledger Fabric 开发的区块链验证器工具验证任何更改。该测试模拟了对账本完整性的直接攻击，而大型科学数据文件本身仍保留在其链下存储位置，不受影响。

4 结果

为了进行完整性验证测试，直接操纵了区块链文件，这代表了一种极端场景，可用于展示数据完整性机制的稳健性。通过直接修改底层账本，该测试绕过了应用层保护，以挑战区块链的核心防护机制。由此产生的哈希不匹配清晰地暴露了篡改行为，展示了区块链如何通过对交易进行密码学链接来确保防篡改可验证性。此外，由于每个区块都相互链接并通过共识进行验证，即使是孤立的更改也会破坏账本的完整性，并能够被可靠地标记出来。

然而，这类篡改在真实环境中不太可能发生。在本实验中，从所有交易中随机选取了十三条数据。区块级验证的结果如图 5 所示，突出显示了链上数据遭到篡改时出现的哈希不匹配。**结果采用 JSON 格式 [10]。

使用**区块链验证器**分析遭篡改的数据时，检测到了几处具体异常，证实数据遭到了篡改。首先，显然，遭篡改的

[10] 结果存储库：https://github.com/GSMier/ProyectoLago/tree/main/resultados

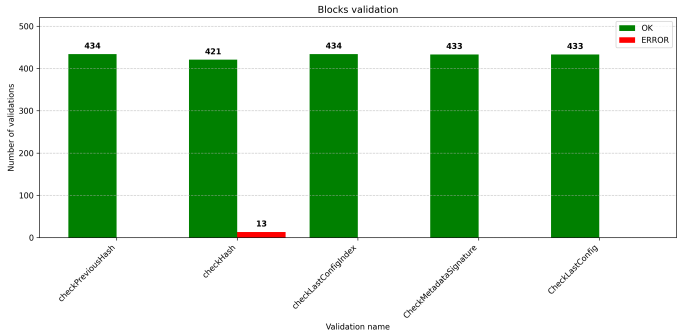


图 5. 区块验证

区块内容的哈希值已与存储在其头部中的原始数据哈希不再匹配，这是数据泄露的主要迹象。该工具还发现了 Hyperledger Fabric 中一种意料之外的行为：修改区块内容后，存储在其自身头部中的区块哈希值并未重新计算；若非其他检查手段，这可能会使篡改检测变得复杂。此外，如图 6 所示，此次篡改导致大量交易签名失效。在本例中，**65 个数字签名**未能通过验证，这证实单笔交易的完整性也遭到了破坏。

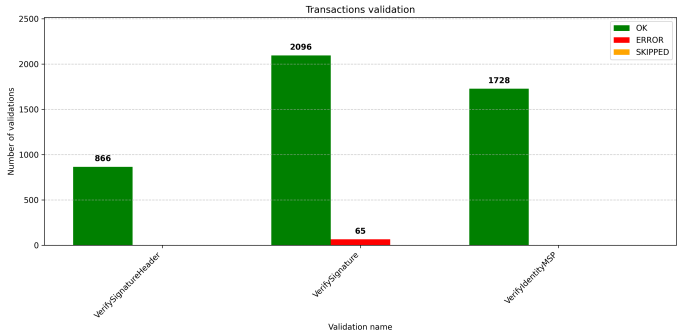


图 6. 交易验证

Hyperledger Fabric中的交易验证高度依赖通过**成员服务提供商 (MSP)**生成的数字签名。MSP 负责向所有授权参与者颁发加密身份。随后，每笔交易都会使用与参与者身份关联的私钥进行签名。这些直接取决于交易负载的数字签名发挥着两项关键作用。首先，它们可提供身份验证，证明交易是由合法的网络成员创建的。其次，它们可确保完整性，因为交易数据在签名后发生任何更改，都会导致签名验证失败。这增添了至关重要的一层安全性，因为每笔交易都包含各自独立的

可验证证明，由其背书者和所有者提供，因此任何篡改都能通过签名不匹配迅速发现。

在本例中，有 65 个签名未通过验证。这是因为每笔交易可能包含一个或多个签名，具体取决于验证所需的通过验证的签名数量。对于 **ScientificDataCollection** 链码，批准一笔交易需要四个签名。需要注意的是，还有其他链码正在执行，例如与新 **链码**实施生命周期相对应的链码。

5 讨论

区块链作为一种去中心化账本技术，通过实现和支持高性能计算（HPC）分布式系统，为大型科学基础设施面临的各种挑战提供了颇具前景的解决方案。区块链改善了整个先进计算连续体中闲置计算资源的利用，涵盖从边缘设备到云计算能力和存储的各个环节。在应用方面，部署区块链以建立去中心化高性能计算市场，并与其他区块链计划保持一致，有助于促进计算能力的可持续性和共享经济[17]。然而，在此背景下，资源和数据的贡献仍是一个颇有争议的话题，需要进一步讨论，才能准确评估这些资源和数据的价值。例如，在成本效益方面，区块链利用未充分使用的资源，使基于区块链的高性能计算平台能够显著降低大规模计算相关开支，相较于传统超级计算机或云计算服务尤为如此。然而，要确保这些优势得以实现，增强信任和安全性至关重要；区块链固有的透明度和不可篡改性可确保计算任务安全执行、数据完整性得到保障，并确保参与者获得补偿。

为防止数据篡改，仍需开展更多工作。鉴于 LAGO 以及全球许多其他高性能计算项目的数据样本规模庞大，将所有数据存储在区块链中并不可行，但记录哈希值及其他元数据应当能够为科学家提供一种检查数据完整性的方式，并验证所有研究团队无论身处何时何地，使用的都是相同的样本。

6 结论与后续工作

针对科学数据完整性开发基于区块链的原型，证明了利用 Hyperledger Fabric 增强大规模分布式数据集安全访问、可追溯性和防篡改可验证性的可行性。将该原型应用于 LAGO 项目公开数据的一个子集后，结果表明，即使科学记录存储在去中心化的异构环境中，区块链仍能有效支持其不可篡改性和可验证性。

区块链基础设施的密码学机制确保能够立即检测数据集中的任何未经授权的改动，从而满足科学研究可复现性和信任的要求。通过确保数据未被篡改，可以避免重新处理大量信息；而在高性能计算（HPC）环境中，这项任务会产生高昂的能源成本。这一优势对于 LAGO 等项目尤为重要，因为此类项目的数据量庞大且生成成本高昂。

在实施过程中，节点同步和一致性方面出现了挑战，尤其是在对等节点失去连接时。这种连接中断直接影响了原型的可靠性，

导致节点的本地账本可能与网络其余部分不一致，形成安全隐患。向此类节点提交交易时，由于数据与哈希值不匹配，会导致错误，从而威胁系统的完整性。我们在应用层解决了这些问题：实施了一种机制，阻止向处于损坏状态的对等节点提交交易；同时集成了重新初始化过程，以便在连接恢复后恢复系统一致性。然而，这些挑战凸显了进一步提升可扩展性、容错性和运维自动化的重要性，以便在完整的生产环境中部署。

依托其他需求技术（如人工智能（AI）），可弥补彼此在可追溯性、互操作性和可扩展性方面的不足。将人工智能与区块链和高性能计算结合，是我们正在开展的工作之一。

致谢。作者感谢 CYTED 共同资助的 LAGO-INDICA 网络（524RT0159- LAGO-INDICA：开放科学数字基础设施）以及桑坦德工业大学（UIS），特别是高级与大规模计算（CAGE）研究小组，为本研究提供机构支持和资源。本文使用了人工智能技术进行校对和润色（OpenAI，2025年；Gemini，2025年）。

参考文献

1. Hirsh, S., Alman, S.W.: 区块链。ALA Neal-Schuman，芝加哥（2020） 2. Silva, G., Klôh, V., Yokoyama, A., Gritz, M., Schulze, B., Ferro, M.: SMCis：面向高性能计算环境的科学应用监控与预测。载于：高性能计算系统，第 69–84 页。Springer International Publishing，查姆（2020） 3. Mantripragada, K., Binotto, A., Tizzei, L.P.: 面向高性能计算环境和云的科学应用自适应自动扩缩方法。arXiv.org（2015） 4 15 5 2025 05 . LAGO 项目主页，https://lagoproject.net/，最后访问日期 / / 。Iván Sidelnik, Hernán Asorey 6. Iván Sidelnik, Hernán Asorey, Nicolás Guarin, Mauricio Suaréz Durán, Fabricio Alcalde Bessia, Luis Ho-racio Arnaldi, Mariano Gómez Berisso, José Lipovetzky, Martín Pérez, Miguel Sofo Haro, Juan Jerónimo Blostein 7. Rubio-Montero, A.J., Pagan-Munoz, R., Mayo-Garcia, R., Pardo-Diaz, A., Sidelnik, I., Asorey, H.: 拉丁美洲巨型天文台（LAGO）标准化模拟的新型云端框架。载于：2021年冬季仿真会议（WSC），第 1–12 页（2021） 8. LAGO 合作组织，*LAGO*数据管理计划（*DMP*）文档 *1.1*，2022 年。 [访问日期：2025- 05-10]。9. CompTIA：区块链实施的七个阶段。 https://www.mspinsights.com/doc/ phases-of-blockchain-implementation-0001，最后访问日期 / / 2025 05 18 10。Raj, P., Deka, G.C.: 区块链技术：平台、工具和用例。Academic Press（Elsevier 旗下品牌），马萨诸塞州剑桥（2018） 11. Hyperledger Fabric v2.5 的新功能， https://hyperledger-fabric.readthedocs.io/en/ release-2 18 12 5 2 5, 2025 05 . /whatsnew.html# what-s-new-in-hyperledger-fabric-v - 最后访问日期 / / 。Coelho, R., Braga, R., David, J.M.N., Stroele, V., Campos, F., Dantas, M.: BlockFlow：用于协作科学实验中建立信任与可追溯性的区块链架构。Journal（多德雷赫特），（2022） 13. Hoopes, R., Hardy, H., Long, M., Dagher, G.G.: SciLedger：基于区块链的科学工作流溯源与数据共享平台。载于：2022年IEEE第八届国际协作与互联网计算会议（CIC），第 125–134 页。IEEE（2022） 14. Zhang, F., Wang, Z., Guo, R., Qu, G.: 地球观测数据溯源：基于区块链的解决方案。*IEEE*工业信息学汇刊 **20**(7)，9548–9556（2024）

14 G.Mier 等人。

15. Xu, H., Zhang, L., Liu, Y., Cao, B.: 存在恶意干扰时基于 RAFT 的无线区块链网络。*IEEE*《无线通信快报》**9**(6)，817–821（2020） 16. Hyperledger Fabric 文档，https://hyperledger-fabric.readthedocs.io/en/latest/ ledger/ledger.html，最后访问日期 2025 14 17 05 2018，16 / / 。Griffin, M.（3月）。企业家认为区块链是云计算和高性能计算的未来。311研究所。